#38

EXHIBIT ONE

order to achieve the proper efficiency of translation. These rules are much less clear for viral RNAs, due to the interfering and overshadowing influences of other processes such as replication and packaging which require particular sequences.

The second manifestation of codon usage preference occurs in the selection of synonymous codons by different tRNA species charged with the same amino acid. Such a population of tRNAs is known as a family of "iso-accepting" tRNAs. As would be expected for strongly expressed genes, there is a clear correlation between the relative amounts of different iso-accepting tRNA species and the use of corresponding codons (Ikemura, 1981). In other words, the higher the concentration of a particular iso-accepting tRNA, the more often the corresponding codon appears in the sequence of the strongly expressed gene; on the other hand, this means that translation may be modulated and controlled by rare codons, for which the corresponding tRNAs occur in trace amounts only. Such codons may be AUA, coding for isoleucine, CUA (leucine), CGG and CGA (arginine), and GGA (glycine). These codons are marked by an arrow in Table 7-4, and, indeed, they are hardly used at all. It is not known whether organisms really use this mechanism to control gene expression.

It should be noted here that the choice of a codon for which there is a limited supply of a corresponding charged tRNA would inevitably cause an imbalance in the tRNA population of an organism. This in turn would not only slow down translation but would also make the system more prone to errors. One may imagine, for example, a competition between correct and false tRNAs at the ribosome A site prepared for the entry of an aminoacylated tRNA. If, due to its low concentration, the correct tRNA were too slow to interact, the false tRNA would associate with the ribosome and, hence, a false amino acid would be incorporated into the growing polypeptide chain. This process may even be associated with alterations of the reading frame if the structure of the false tRNA prevents the proper entry of the next tRNA, and this has, indeed, been observed for several suppressor tRNAs. It is the basis of the phenomenon known as frameshift suppression. Suboptimal translation conditions of this kind have been artificially induced *in vivo* by starving bacterial cells for certain amino acids or *in vitro* by the addition of certain tRNAs to cell-free systems (Roth, 1981; Weiss und Gallant, 1983).

The significance of an appropriate codon choice for the expression of foreign genes in heterologous organisms has never been convincingly documented; nevertheless, especially since other unknown parameters may affect heterologous gene expression, the rules mentioned in this section should be followed as closely as possible in order to approach natural conditions. For chemically synthesised genes, for example, codons should be selected in accordance with the frequencies with which such codons occur in the desired host organism (*cf.* Section 11.2.2.1).

7.4 Construction of Expression Vectors

Several strategies using regulatory sequences discussed in the preceding sections have been pursued to optimise the expression of genes. In principle, these strategies are aimed at the construction of vectors allowing the synthesis either of fusion proteins comprising vector and insertion sequences (Fig. 7-44A) or of pure proteins exclusively encoded by the insertion (Fig. 7-44B). The first construction is referred to as a translational fusion, the second as a transcriptional fusion. The following selected examples will clarify this distinction.

7.4.1 Synthesis of Fusion Proteins

In order to obtain a hybrid protein, the foreign DNA must be inserted into an expressable vector gene in such a way that the reading frame in this

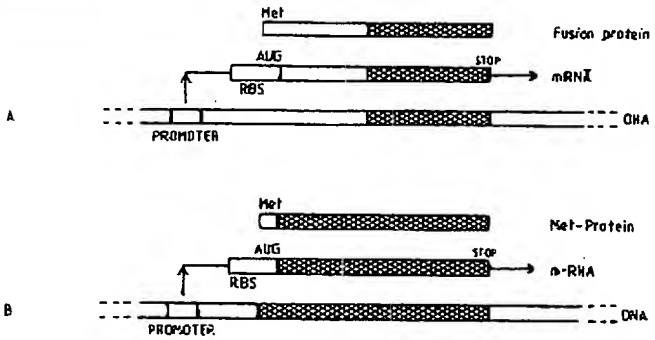


Fig. 7-44. Construction of expression vectors.
Two approaches are shown, namely the formation of fusion proteins (A), and the formation of native proteins (B) from recombinant DNA. RBS signifies a ribosomal binding site. Met-protein indicates that proteins obtained from recombinant DNA by approach (B) always carry an *N*-terminal methionine residue. Bacterial sequences are represented as open, eukaryotic sequences as hatched bars.

gene is conserved. The synthesis of hybrid mRNA is initiated by the prokaryotic promoter and its translation is controlled by the corresponding ribosome binding site. The first practical application of fusion proteins allowed the expression of rat insulin, rat growth hormone, and human growth hormone, and demonstrated for the first time that bacteria are, indeed, capable of expressing eukaryotic coding sequences.

7.4.1.1 Expression of Rat Insulin

The starting point in this case was the insertion of a rat insulin cDNA into the *Pst* I site of pBR322 by homopolymeric poly(dG)-poly(dC) tailing (Villa-Komaroff *et al.*, 1978). The variable lengths of these tails guaranteed that at least one in three clones contained the right reading frame; however, since the cDNA could be inserted in two

Howard M. Goodman
Boston

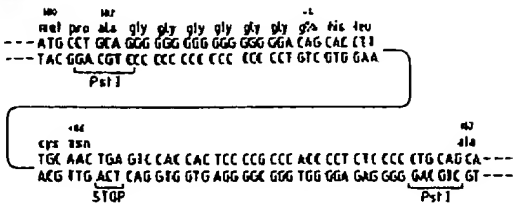


Fig. 7-45. Insertion of the rat insulin DNA sequence into the *Pst* I site of the β-lactamase gene of pBR3
The two *Pst* I sites, and amino acids 182 and 183 of β-lactamase, which are separated by the insertion, are pr
bold-face. The insulin insertion begins with amino acid Gln (position +4) in the B chain, and ends with aspar
of the proinsulin. The order of the insulin peptides is pre-B-C-A. (Villa-Komaroff *et al.*, 1978).

different orientations, only one sixth of the clones containing the desired insulin insertion would also make insulin. In spite of these obvious limitations, cloning by homopolymeric tails was the method of choice because the exact sequence of the cDNA was not known and the desired constructions therefore, could not be planned in advance (*cf.* also Section 3.2). The structure of one the rat insulin clones is shown in Fig. 7-45. Starting with position 182 (ala) the sequence of the β-lactamase gene then proceeds with polyglycine and eventually reaches the insulin sequence at amino acid "+4" (gln) of proinsulin. The desired fusion protein was detected by immunological techniques (see Section 11.2.3.2).

7.4.1.2 Expression of Rat Growth Hormone and the Structural Protein VP1 of Foot and Mouth Disease Virus

A much more direct strategy was pursued for the construction of vectors coding for rat growth hormone (Seeburg *et al.*, 1978). The rat growth hormone cDNA possesses a single *Pst* I site at position "-24" of the prepeptide region, which allowed it to be annealed with the *Pst* I site of the β-lactamase gene of pBR322 in such a way that the reading frame was conserved (Fig. 7-46); in addition, the strategy employed for the construction of this expression vector also allowed a selection for clones containing the desi binant.

The starting material was a cDNA c the *Hind* III site of pBR322. The inse excised and recloned into the *Hind*III plasmid pMB9 which lacks the β-lactamase The resulting plasmid, pMB9-RGH, exp only low levels of tetracycline resistance sin *Hind*III site is located within the tetrac promoter region (Fig. 7-47; *cf.* also Fig. 4. Tetracycline resistance was restored an coding sequence of the rat growth hormon brought under the control of the β-lact promoter by replacing a small *'PstI-BamH* ment of pMB9-RGH with a correspondin ment of pBR322 (*cf.* also Fig. 7-46). This e sion vector could be distinguished from p RGH by its increased tetracycline resistan μg/ml instead of 5 μg/ml). The hybrid gene expression vector coded for a chimaeric p

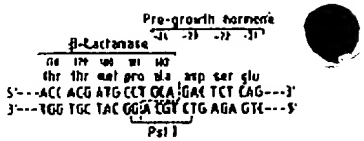


Fig. 7-46. DNA sequence in the vicinity of the *P* of a hybrid vector containing a fusion of the β-lac gene with the gene for rat growth hormone.

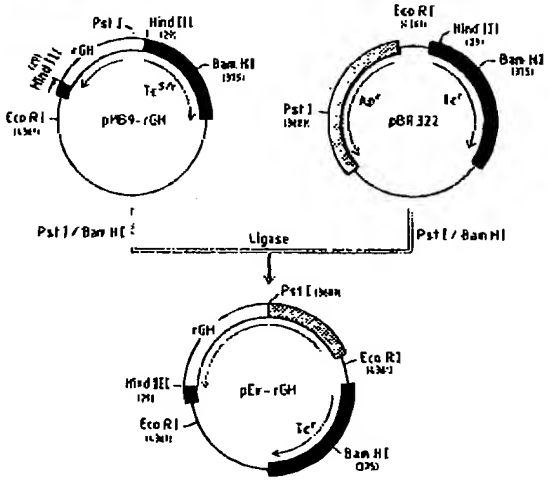


Fig. 7-47. Expression vector for rat growth hormone (rGH).
The expression plasmid pEx-rOH is constructed by replacing the small *Pst*I-*Bam*HI fragment of plasmid pMB9-rGH by the smaller *Pst*I-*Bam*HI DNA fragment of pBR322. Numbers in brackets refer to co-ordinates in plasmid pBR322. Arrows indicate the direction of transcription or translation. In pMB9-rGH, the promoter of the tetracycline resistance region is interrupted by the rGH insertion; expression of tetracycline resistance is therefore markedly reduced. Transformants are resistant to 5 µg tetracycline/ml, while full expression in pBR322 or the expression vector pEx-RGH allows selections with 20 µg/ml. Tc^r regions are represented as black, Ap^r regions as hatched, and rGH regions as open bars. (Seeburg *et al.*, 1978).

395 amino acids in length. It comprises 181 *N*-terminal amino acids derived from the β-lactamase gene and 214 *C*-terminal amino acids of the pre-growth hormone and was, indeed, detected as a protein with a molecular weight of 46 000 in a mini-cell test system (see Section 11.2.3.3); however, the amount of hybrid protein was only one-fifth the amount of β-lactamase produced by pBR322.

Another example is the synthesis of a fusion protein containing a part of protein VP1 of Foot and Mouth Disease Virus (Fig. 7-48; Küpper *et al.*, 1981). In this case, cloning started with the insertion, into vector pPLc24 cut with *Bam*HI and *Hind*III, of an 849 bp *Bam*HI-*Hind*III fragment coding for amino acids 9 to 292 of the desired protein (Fig. 7-28). The fusion protein obtained was 395 amino acids in length and consisted of 98 N-terminal amino acids of MS2 replicase, 284 amino acids of the desired viral protein, and thirteen plasmid-derived amino acids added because of read-through into neighbouring vector sequences.

7.4.1.3 Expression of Human Growth Hormone

Suitable restriction sites are rarely positioned such that they are located at the beginning of a structural gene and also allow this gene to be inserted into the vector gene in the correct reading frame. Quite frequently it is necessary to design special constructions in which linker molecules play an important role. The following

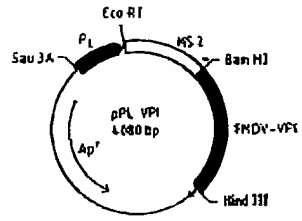
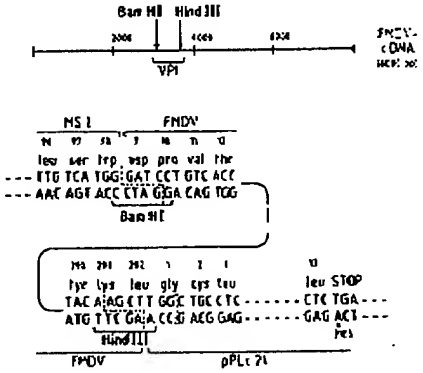


Fig. 7-48. Structure of an expression vector structural protein VP1 of Foot-and-Mouth Virus (FMDV).
The expression plasmid pPLVP1 is derived fr mid pPLc24 which contains the *N*-terminal part replicase under the control of λ promoter P_L (Fi The structure of the FMDV cDNA and the po the VP1 structural protein with flanking *Bam Hind* III sites is indicated in the centre. Show bottom are the sequences around the *Bam Hind* III site in expression vector pPLVP1. T codon is at position 2105 of the pBR322 sequ



example of a human growth hormon (hGH) expressed under the control of promoter may illustrate the point (Martia 1979). The starting material in this case wa ptrpED5-1 (Fig. 7-16) with a *Hind* III site the codon for amino acid 92 of the this *Hind* III site had been joined fragment flanked by a *Hind* III site in untranslated region of the cDNA for growth hormone, the correct reading frame have been lost (Fig. 7-49). It was th necessary to manipulate the *Hind* III ptrpED5-1 to shift the reading frame by o in the recombinant molecule. This was plished by filling-in the 5' protruding en Klenow fragment of *E. coli* DNA polym and adding a synthetic DNA decamer contained a *Hind* III site. As shown in Fi cleavage of the new plasmid ptrpED5 *Hind* III and subsequent ligation with *Hind* hGH cDNA conserved the correct reading

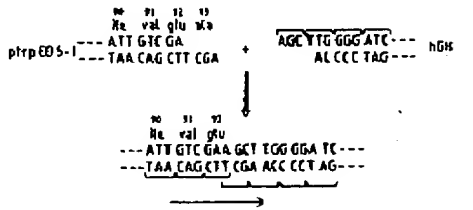


Fig. 7-49. Fusion of a *Hind* III site in plasmid ptrpED5-1 with a *Hind* III-flanked cDNA fragment o human growth hormone (hGH).
Brackets indicate the reading frames, the arrow denotes the direction of translation. The hGH section derived from the 5' untranslated region.

in the recombinant molecule. After induction of transformed bacteria with 3-β-indolylacrylic acid, the expected fusion protein with a molecular weight of 34 kDa (322 amino acids) was obtained with a yield corresponding to 3% of the total cellular protein. This value, however, is only one sixth of that expected from induction of the non-recombinant plasmids ptrpED50 or ptrpED5-1.

7.4.1.4 Expression of Somatostatin

Somatostatin is a peptide hormone consisting of fourteen amino acids, which controls the secretion of a number of other hormones, such as insulin and glucagon. The starting material for cloning and expression of a somatostatin fusion protein was plasmid pBR322 and a synthetic gene with the coding sequence of somatostatin (Itakura *et al.*, 1977). The synthetic gene was pieced together by annealing eight different oligonucleotides (A to H) and contained terminal protruding *Eco* RI and *Bam* HI ends, respectively. As shown in Fig. 7-51, an additional methionine codon was introduced directly in front of the *N*-terminal alanine codon; the carboxy-terminal cysteine codon is followed by two stop codons.

Since the somatostatin gene was to be expressed under the control of the *lac* system, the first step was to generate a suitable vector containing the *lac* regulatory region. This was accomplished

Axel Ullrich
San Francisco

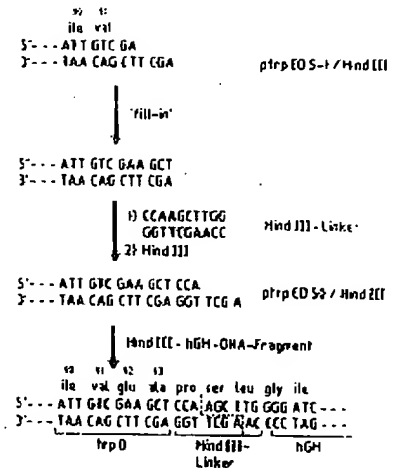


Fig. 7-50. Construction of vector ptrpED50 and cloning of the gene for human growth hormone (hGH). In order to move the reading frame at the *Hind* III site of ptrpED5-1 by one base, the 5' ends were filled-in and fused with a decameric *Hind* III linker. The fusion peptide contains the first 93 amino acids of the *trpD* protein, three amino acids encoded by the *Hind* III linker, nine amino acids of the 5' untranslated region, and 217 amino acids of pre-hGH (total of 322 amino acids). (Martial *et al.*, 1978).

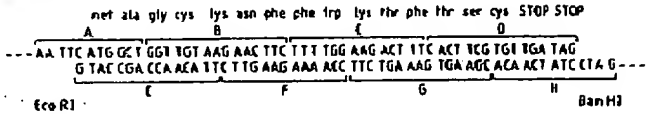


Fig. 7-51. A chemically synthesised somatostatin gene.
Shown are the eight oligonucleotides, A-H, used as building blocks, the *Eco* RI and *Bam* HI termini, the ATG codons, the start codon, and the two stop codons.

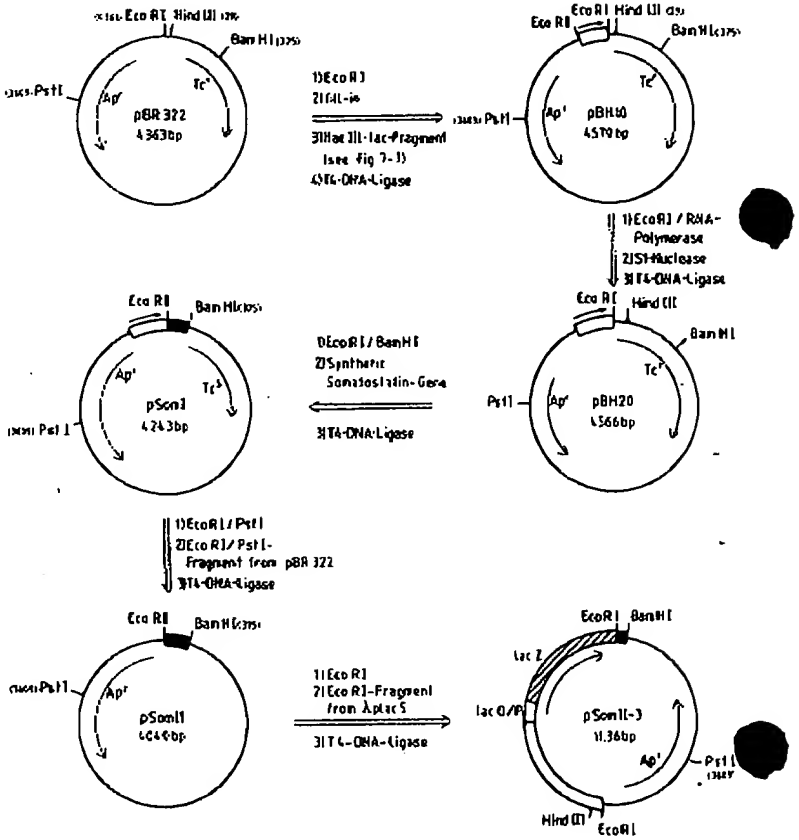


Fig. 7-52. Construction of an expression vector for the hormone somatostatin.
Open bars represent the *lac* control region; hatched bars the *lacZ* gene and black bars the somatostatin gene indicate the direction of transcription.

by partially digesting λp*lac*5 DNA with *Hae* III. The resulting mixture of DNA fragments contained a 203 bp *Hae* III fragment coding for the entire *lac* control region and the first seven amino acids of β-galactosidase (*cf.* Fig. 7-7). The mixture of DNA fragments was then ligated with pBR322 DNA which had been linearised with *Eco* RI and subsequently filled-in to convert its protruding 5' ends to flush ends. Ligation of filled-in *Eco* RI ends with blunt *Hae* III ends generated new *Eco* RI ends in the recombinant molecules at the sites of fusion (see also Section 2.1.2.1).

Transformants obtained from this DNA mixture containing the desired *Hae* III fragment were identified as blue colonies on agar plates containing Xgal (Fig. 7-6). This screen did not distinguish between the two possible orientations of the *Hae* III fragment; however, since there was an asymmetrically positioned *Hha* I site directly following the stop codon of the *lacI* gene on the *Hae* III fragment (Fig. 7-7), it was easy to determine the orientation of the inserted *Hae* III fragment. The desired orientation was found in vector pBH10 (Fig. 7-52), in which *lac* transcription proceeds toward the tetracycline resistance region.

An unusual procedure was used to selectively remove the distal *Eco* RI site in pBH10. *E. coli* RNA polymerase binds to promoter regions in the absence of nucleoside triphosphates. In pBH10 binding occurs at the *lac* promoter and also at the tetracycline promoter region, 20 bp away from the proximal *Eco* RI site (Fig. 7-53). A region spanning approximately 20 bp around a Pribnow-Schaller box is protected by RNA polymerase, blocking further enzymatic attack. Therefore the proximal, but not the distal *Eco* RI site in pBH10 was protected, and only the distal recognition site was cleaved by *Eco* RI in the presence of RNA polymerase. Subsequent digestion with nuclease S1 was used to convert this *Eco* RI terminus into a blunt end. Ligase treatment yielded plasmid pBH20 with only one *Eco* RI site (Fig. 7-52).

Plasmid pBH20 was prepared for the incorporation of the synthetic somatostatin gene by first digesting it with *Eco* RI and *Bam* HI and separating the resulting two fragments by gel electrophoresis. The larger of the two fragments was treated with phosphatase and annealed with the synthetic somatostatin gene. Transformants were selected for ampicillin resistance and screened for tetracycline sensitivity. Fig. 7-54A shows the DNA sequence from the region of the insertion in clone pSomI. This clone should yield a fusion protein of 24 amino acids, the expression of which is controlled by the ribosome binding site of the *lacZ* gene. Suitably induced bacteria were then treated with cyanogen bromide in order to cleave somatostatin from the entire mixture of proteins. Since cyanogen bromide cleaves peptides specifically at the carboxyl group of methionine residues

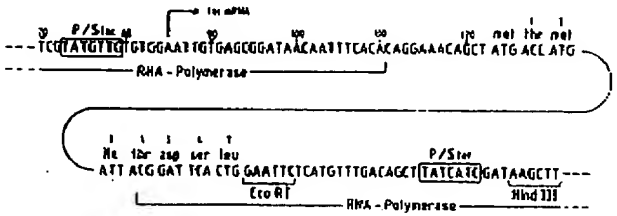
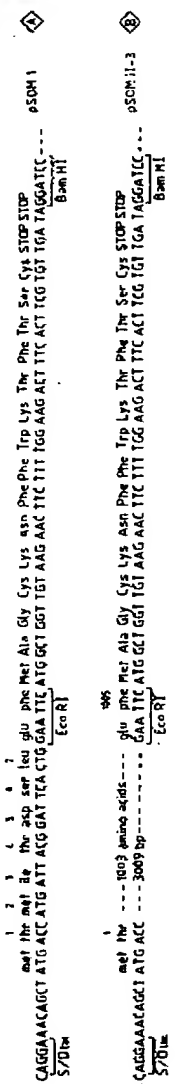


Fig. 7-53. DNA sequences between the *lac* and *tet* promoter regions in plasmid pBH10.
The Pribnow-Schaller boxes are framed, the proximal *Eco* RI site of the *tet* promoter and the *Hind* III site are bracketed. Binding sites for RNA polymerase, which extend approximately 35 bp to the left and right of the Pribnow-Schaller boxes, are indicated by brackets. It is apparent that the proximal *Eco* RI site lies in a region protected by RNA polymerase. For the numbering in the *lac* region see legend to Fig. 7-7.



this procedure should have yielded functi somatostatin; nevertheless, all attempts to d the hormone in various bacterial extracts unsuccessful. These negative results prob resulted from proteolytic cleavage of the horm (*cf.* Section 7.5).

A new plasmid, pSomII, was constructed i hope that the presence of a large peptide w prevent this proteolytic attack (Fig. 7-52). In plasmid the smaller *Eco* RI-*Pst* I fragment the *lac* region of pSomI was replaced corresponding *Eco* RI-*Pst* I fragment of pBR Transformants were selected for ampicillin r ance and screened on Xgal plates for the abs of *lac* operator DNA. A *lac* region containin entire control region and the codons for 1 00 of 1 021 amino acids of β-galactosidase, than the first seven amino acids, was to replace the missing *lac* region. *Eco* RI fragment of 7.45 kb was obtained λp*lac*5 DNA. As shown in Fig. 7-54B, the co reading frame was retained in this constru and a fusion protein of 1 020 amino acids wit amino acid sequence of somatostatin at C-terminal end was obtained. When total cel proteins of suitably induced bacteria were tr with cyanogen bromide, somatostatin activit indeed detectable. The yield in uninduced was estimated to be on the order of 0.001-0. of the total protein. This low yield reflects th basal level of transcription from a fully repr *lac* promoter. Induction with IPTG led to a t to sevenfold increase in somatostatin yields. induction experiment confirmed the sequence data indicating that in pSomII-synthesis of somatostatin was regulated by t control sequences. However, the induction was approximately tenfold lower than had

Fig. 7-54. Nucleotide sequence of *Lac* fusion genes.
(A) shows a fusion with only seven *N*-terminal acids of β-galactosidase, (B) a fusion with 1003 acids of β-galactosidase. Both fusions contai additional amino acids derived from the *Eco* I (Itakura *et al.*, 1977).

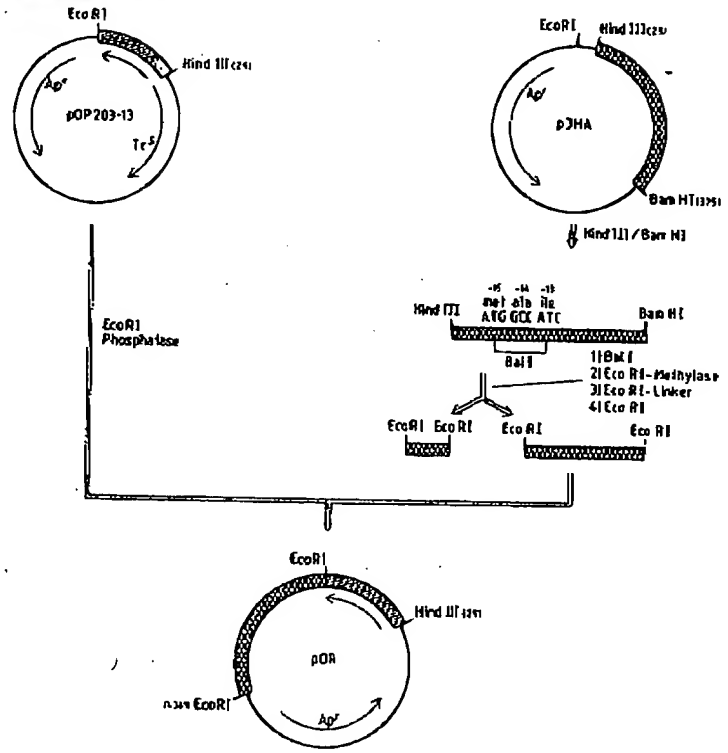


Fig. 7-55. Construction of a plasmid for the expression of the haemagglutinin (HA) gene of human influenza virus type A/Japan/305/57 (subtype H2).
The modifications of the HA DNA fragment which is obtained from pJHA by *Hind* III-*Bam* HI digestion are described in detail in the text. *Lac* control regions are indicated by a stippled bar, HA DNA sequences by a cross-hatched bar (Heiland and Gething, 1981).

expected. Similar observations subsequently have been made with other expression plasmids based on *lac* control elements. There are several possible explanations for this phenomenon, including the selective cleavage of the foreign protein by bacterial proteases, insufficient solubility of the fusion protein during cyanogen bromide cleavage and the instability of the recombinant plasmid.

7.4.1.5 Construction of Expression Plasmids for Influenza Virus Specific Sequences

This case deals with the expression of a DNA copy of an RNA fragment coding for the haemagglutinin (HA) protein of human influenza virus strain A/Japan/305/57 (subtype H2) (Heiland and

Gething, 1981). The vector used was plasmid pOP203-13 (Fig. 7-9) which contains, between the *Eco* RI and *Hind* III sites of pBR322, the same 203 bp of the *lac* control region as pHB10 (Fig. 7-52). The direction of *lac* transcription is anticlockwise, *i.e.*, in the direction of the β-lactamase gene (Fuller, 1982), which means that any DNA inserted into the single *Eco* RI site of this plasmid will be controlled by the *lac* promoter.

The haemagglutinin gene to be expressed was inserted between the *Hind* III and *Bam* HI sites of pBR322 in plasmid pJHA (Fig. 7-55). It codes for the entire 560 amino acids of the haemagglutinin protein and eleven nucleotides of the 5' untranslated region. This sequence must be modified before it can be inserted into the *Eco* RI site of pOP203-13. A *Bal* I site comprising the ATG start codon of the HA gene is important. The DNA fragment obtained by *Hind* III and *Bam* HI digestion is further cleaved by *Bal* I treatment to yield two sub-fragments. The mixture of fragments is first treated with *Eco* RI methylase in order to methylate internal *Eco* RI sites and to render them resistant to *Eco* RI digestion. *Eco* RI linkers are then added by ligation. Following *Eco* RI digestion the sub-fragments are separated from each other and the larger fragment is inserted into the *Eco* RI site of the expression vector pOP203-13 (Fig. 7-55).

Cloning of the large *Bal* I fragment was expected to yield a fusion protein with the structure shown in Fig. 7-56, containing seven *N*-terminal amino acids derived from β-galactosidase, three amino acids coded for by the linker, and 560 amino acids of the haemagglutinin gene. Two *N*-terminal amino acids of the leader sequence of the HA gene were removed by this cloning procedure. Again, the two possible orientations for the inserted gene could be easily distinguished by suitable digestions. Three of the clones obtained expressed antigenic determinants of haemagglutinin, as shown by solid phase radioimmunoassay. The nucleotide sequences of all three clones confirmed that they preserved the correct reading frame, but also showed that they did not have the expected sequence of the hypothetical pOR (Fig. 7-56) at the site of fusion. For unl reasons fifteen amino acids of the signal p and the first ten to fifteen amino acids mature protein were missing. Perhaps the e otic hydrophobic signal sequences were n erated by the *E. coli* host organism i particular case. By way of contrast, other l phobic signal sequences, such as that of l preproinsulin, have been found to be quite in *E. coli* (Chan *et al.*, 1981).

7.4.1.6 General Technique for the Constru of Expression Vectors for Fusion Proteins

In the examples discussed so far, a res within the region of the ATG start always positioned in such a way that the H into the *lacZ* gene (or another suitable either conserved the correct reading fra could be easily arranged to fit into frame. case of the somatostatin gene this was accon ed by suitably planning the chemical syn while it was mere coincidence in the case HA gene. In most cases, however, a conv restriction site will not be available; the foll procedure is therefore recommended for c and expression: the DNA to be expresse example a cDNA, is cut out and isolated f parent plasmid. The example shown in Fig uses a *Pst* I digestion. The next step pl suitable restriction site in the vicinity of the start codon. The DNA is first treated v combination of the enzymes *Exo* III and S1 c *Bal*31 (*cf.* also Fig. 2.1-9). Digestion cond depend on the distance between the c cleavage site (*Pst* I in our example) and the start codon, and must be determine for every individual case. In our exam linkers are then added to the fragment, so can be inserted into the *Eco* RI site of a s vector. Before the fragment containing the H is cloned, it is digested with *Eco* RI and a restriction enzyme which cleaves at site X

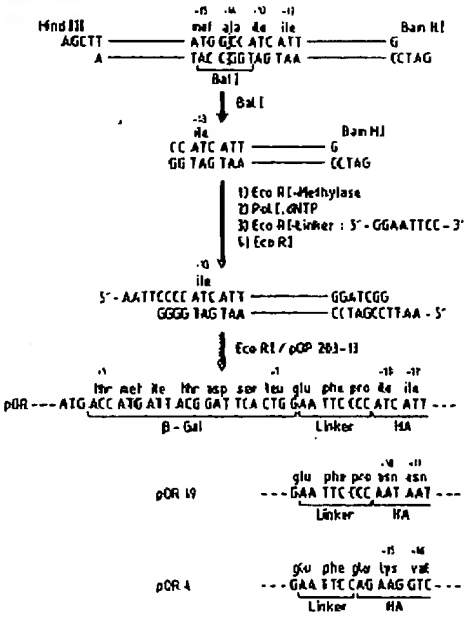
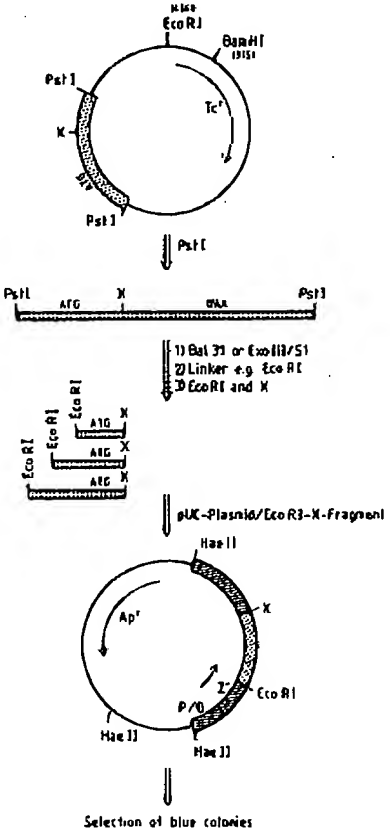


Fig. 7-56. Construction of an expression plasmid by linker technology.
A *Hind* III-*Bam* HI fragment of plasmid pHA (Fig. 7-55) is cleaved into two fragments by digestion with *Bal* I. Only the flanking sequences of the larger of the two fragments are shown. Following an *Eco* RI methylase treatment, the protruding *Bam* HI site is filled-in in order to allow the subsequent addition of *Eco* RI linkers. A further *Eco* RI digestion only attacks the *Eco* RI sites within the linker, but not the internal methylated *Eco* RI site. This DNA fragment is cloned into the *Eco* RI site of plasmid pOP203-13 (Fig. 7-9). Shown is the expected structure of the fusion protein (pOR) consisting of seven amino acids of β-galactosidase, three amino acids encoded by the linker, and two amino acids out of a total of 560 from the haemagglutinin. The actual experiment did not yield clones with the expected structure; instead, plasmids were obtained, which begin with sequences of the mature HA protein and which do not contain the hydrophobic leader sequence. Part of the structures of two of these plasmids, pOR19 and pOR4, are also shown.

the gene to be cloned. This yields a defined right-hand molecular end which can be used at a later stage to reconstruct the entire gene. Although the left end of the fragment is defined by an *Eco* RI site, the distance between this site and the ATG start codon varies in different molecules. Ligation with a suitable vector will therefore yield a wide range of different clones with varying distances between start codon and *Eco* RI site. In addition, the insertions may not be in the correct reading frame. Those clones in the mixture which contain the correct reading frame, can be identified by exploiting the phenomenon of α-complementation (*cf.* Section 2.4.2.3). As in the case of M13 cloning (Section 2.4.2), a number of plasmids, known as pUC plasmids, have been developed for this purpose (Fig. 7-58). These plasmids contain the *lac* regulatory region and a part of the *lacZ* gene which codes for the 59 *N*-terminal amino acids of β-galactosidase (Vieira

and Messing, 1982). The corresponding host strain (JM83) carries the deletion *M15* of the *lac* operon, which removes amino acids 11-41 of β-galactosidase, but retains the entire *C*-terminal part of the enzyme. Each incomplete *lacZ* gene will direct the synthesis of an inactive polypeptide. Together, these polypeptides will be capable of complementing each other by forming aggregates. The resulting enzymatic activity can be detected on Xgal indicator plates as described above (Fig. 7-6).



Plasmids of the type pUC7, 8, 9, 12, 13, 1
19 contain polylinkers within the region
lacZ gene corresponding to the 59 *N*-te
amino acids. These polylinkers allow cloni
DNA fragments with a variety of different e
is interesting to note that the inserted polyli
(and other insertions) do not interfere
α-complementation, as long as they presen
correct reading frame.

When the cDNA fragments with differe
ends described above (Fig. 7-57) are inserte
a polylinker of a pUC plasmid, only those
with insertions in the correct reading fram
β-galactosidase will yield blue colonies. Sin
maximum length of an insertion which still
α-complementation is not known, it is adv
to clone only relatively small DNA fragm
The intensity of the blue colour of diff
is usually quite variable. Clones showing
intense colouration are those giving the hi
levels of expression, and can be used as reci
of the missing part of the gene, in order to
the entire fusion protein.

Fig. 7-57. General approach for the construct
expression vectors directing the synthesis of
proteins.
The starting material can be a cDNA clone
insertion has been removed from the plasmid (b
digestion in this example). By treatment with e
lease *Bal*31 or a combined ExoIII/S1 nuclease dig
a restriction site is positioned close to the start
The example of *Eco* RI linkers shown here may
sitate the use of *Eco* RI methylase if the DNA fra
contains an internal *Eco* RI site. The next step
Eco RI digestion followed by digestion with a
restriction enzyme (X) which should preferab
asymmetrically. The mixture of DNA fragments o
ed is then cloned into a pUC vector (*cf.* also Fig.
Of course, the cleavage site of endonuclease X m
present within the polylinker of the pUC plasmid.
a wide spectrum of pUC plasmids is availa
not be difficult to find a suitable vector
desired cleavage site. Once a suitable clone is iden
site X can be used for the insertion of the
portion of the gene in question, which can be ob
from the original cDNA clone. The *lacO/P* and
regions are represented by hatched bars; the in
by stippled bars.

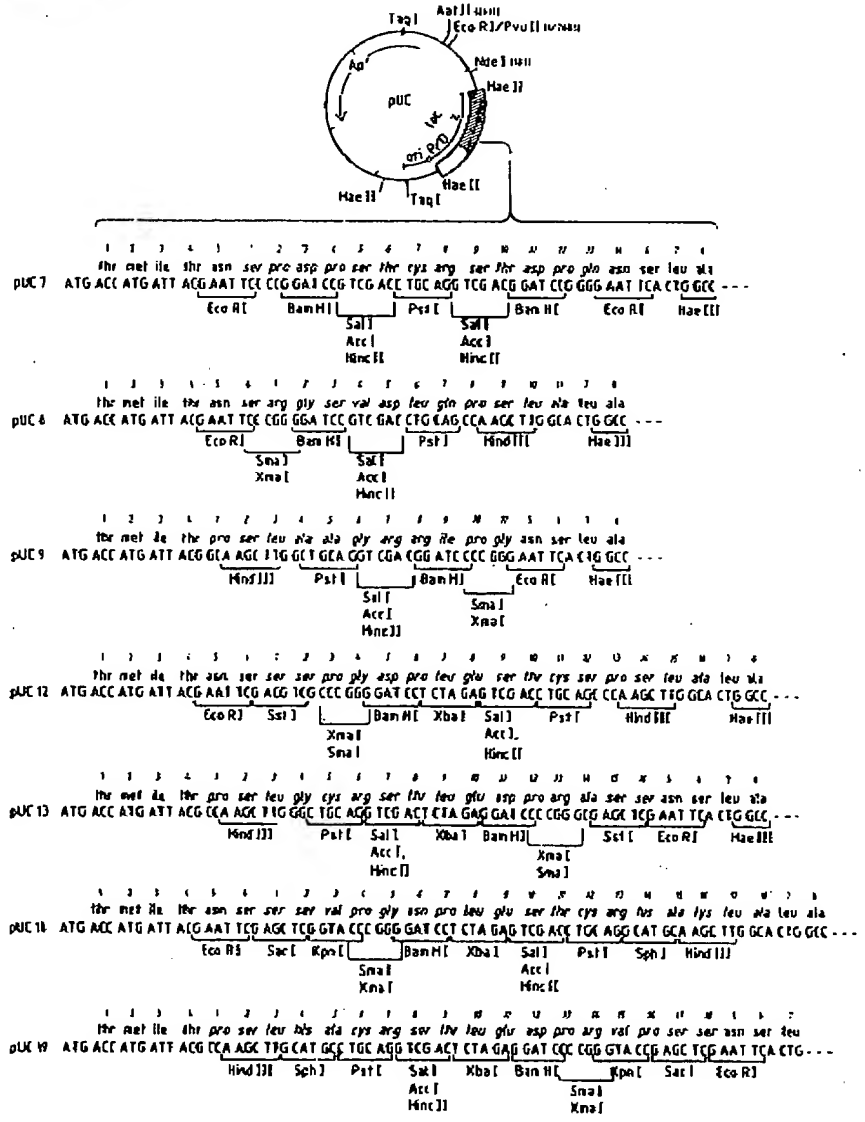


Fig. 7-58. Structure of pUC plasmids.
pUC plasmids (*cf.* Fig. 2.4-22) are derived from the 2297 bp *Pvu* II-*Eco* RI fragment of pBR322, which contains the origin of DNA replication (*ori*) and the coding region of β-lactamase (Ap^r). *Pst* I, *Hind* III and *Acc* I sites were removed by mutagenesis. pUC plasmids carry a 433 bp *Hae* II fragment with *lac* control elements (*lac* promoter (P) and operator (O); open bars) inserted into the *Hae* II site in the immediate vicinity of the replication origin at position 2352; in addition, they contain the coding region for a functional β-galactosidase α-peptide (*lacZ*) (hatched bar). Short polylinker regions within this regions provide multiple recognition sites for various restriction endonucleases. Amino acids encoded by polylinker insertions are printed in italics. Numbers in parenthesis are pUC18 co-ordinates (Appendix B-4; Vieira and Messing, 1982; Yanisch-Perron *et al.*, 1985).

In summary, the successful synthesis of hybrid proteins with prokaryotic and eukaryotic components has been described in detail for several systems. This approach has considerable advantages, particularly the fact that fusion proteins with the large, 1 000 amino acids-long *N* terminus from β-galactosidase often are insoluble within the bacterial cell. Such fusion proteins thus are protected from proteolytic degradation (see below) and are easily purified; however, it should be kept in mind that this strategy also has its limitations. There is no doubt that it permits the detection of antigenic determinants in the fusion protein. Actually, it allowed the initial demonstration of the possibility of expressing eukaryotic DNA sequences in prokaryotes; however, if the eukaryotic proteins in question are to be obtained in a pure form, the original protein must be separable from the bacterial component of the chimaeric fusion products. Cyanogen bromide cleavage, which was used in the case of somatostatin, is restricted to proteins, such as some proinsulins, which do not contain internal methionine residues. In other cases enzymatic cleavage must be employed. Since the codons for suitable specific amino acids (for example arg and lys residues for tryptic cleavage) usually are not found in desired positions on the vector, this opens important vistas for linker technolog strategy is similar to that described above 7-57). Instead of *Eco* RI linkers, more co linkers, coding for protease-sensitive amino would have to be employed.

Even in these cases difficulties may be rienced if internal protease-sensitive amino occur somewhere along the polypeptide ch the fusion protein. Although it will depend l on the individual case, it can be safely said th synthesis of fusion proteins will be partic useful for the production of small protein peptides. A very good example is the prod of endogenous opiate peptides (*cf.* Obsuye 1983).

7.4.2 *Synthesis of Unique Protein Bacteria*

In contrast to procedures described above process, known as transcriptional fusion, directly at the production of unique, non-polypeptides. In this case protein synthesis s not initiate from the first methionine residu prokaryotic leader peptide, such as *lacZ* or but from the first methionine of the d polypeptide itself (Fig. 7-44B). Biologically constructions, therefore, usually contain an i ible prokaryotic promoter and a hybrid ribo al binding site consisting of a bacterial S Dalgarno sequence and a correctly spaced codon which does not have to be of bac origin. In the case of a eukaryotic protei ATG may correspond to the initiation cod the eukaryotic gene itself. The individua ments of such constructions will be describe the *lac*, *trp* and λ systems as examples

7.4.2.1 *The lac System*

M. Ptashne and co-workers have develope concept of a portable *lac* promoter which c placed at a suitable distance in front of a de

structural gene to allow the synthesis of the corresponding gene product as a pure unfused protein. The *lac* operon contains a suitable *Alu* I site between the ribosomal binding site and the start codon of the *lacZ* gene. A 95 bp long *Alu* I fragment containing almost the entire *lac* promoter region, the initiation site for mRNA synthesis, the S/D sequence, and five additional base pairs therefore can be isolated from chromosomal DNA (Fig. 7-7). This fragment also can be obtained from plasmid pGL101, in which this fragment is flanked by an *Eco* RI and a *Pvu* II site (Fig. 7-10) (Lauer *et al.*, 1981). Since *Pvu* II (CAG/CTG) recognises a hexanucleotide sequence comprising the *Alu* I recognition sequence (AG/CT) and produces blunt ends at the same position as *Alu* I, *Eco* RI/*Pvu* II cleavage of pGL101 yields a DNA fragment with the desired blunt *Alu* I ends immediately downstream from the *lac* S/D sequence. This fragment contains an S/D sequence, but lacks an ATG codon, and must be placed at the proper distance upstream from a structural gene. For this purpose, the gene in question should preferably contain a unique

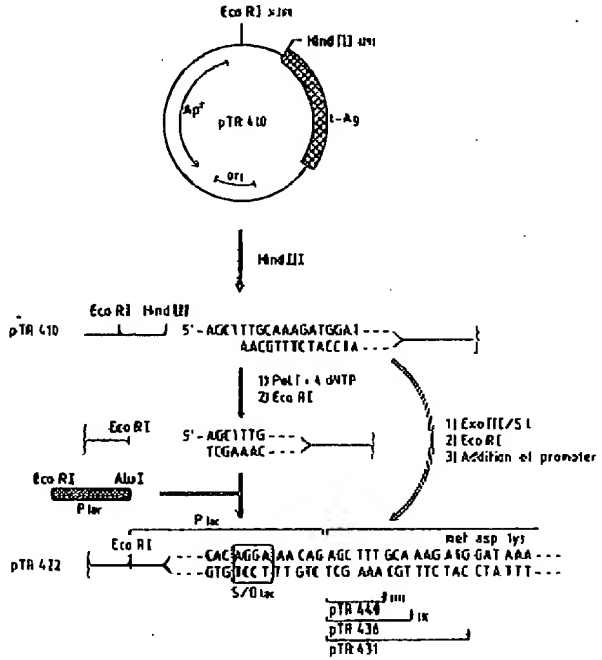


Fig. 7-59. Expression plasmid for SV40 t-antigen.
Vector pTR410 consists of parts of pBR322 comprising the coding region for β-lactamase (Ap^r) and the origin of DNA replication; in addition, it contains SV40 sequences with the entire coding region for the small t-antigen (cross-hatched bar). Manipulations near a *Hind* III site upstream of the start codon for t-antigen are described in detail in the text. P_{lac} signifies the portable *lac* promoter fragment (stippled bar; see also text). Small numbers at the arrow heads of deletions pTR440 and pTR436 indicate the distances between the S/D*lac* regions and the start codon.

recognition sequence immediately 5' of its ATG codon, which is, of course, rare. In the case of small t-antigen of SV40, however, the *Hind* III site in the SV40 *Hind* III-B fragment (which is 1 169 bp in length), is only twelve bp upstream of the ATG start codon of t-antigen. When this DNA fragment with filled-in *Hind* III ends was annealed with a portable promoter, the construction designated pTR422, shown in Fig. 7-59, was obtained (Roberts *et al.*, 1979b).

In order to obtain clones in which the distance between the ATG codon and the *lac* S/D sequence is shortened, the DNA first was digested with *Hind* III and then subjected to a partial digestion with exonuclease III. Blunt ends were generated by S1 nuclease treatment and the molecule was circularised after *Eco* RI cleavage and addition of the portable promoter. This procedure yielded inserts of variable length due to the unspe of the exonuclease III reaction. The clone then screened for expression of the t protein. In this examp clone pTR436, in the distance between the S/D sequence a ATG start codon was 8 bp, was particularly (Fig. 7-59). Plasmid pTR440 is only weakly and the starting plasmid pTR422 and a de pTR431, were completely inactive.

In a similar case the starting materials w same as those described above, name portable *lacUV5* promoter, and the *Hind* III-B fragment with the coding regi the small t-antigen (Thummel *et al.*, 1981 strategy employed, however, differed fro described above in that a *Hind* III link introduced between the S/D sequen ATG start codon (Fig. 7-60). This *Hi*

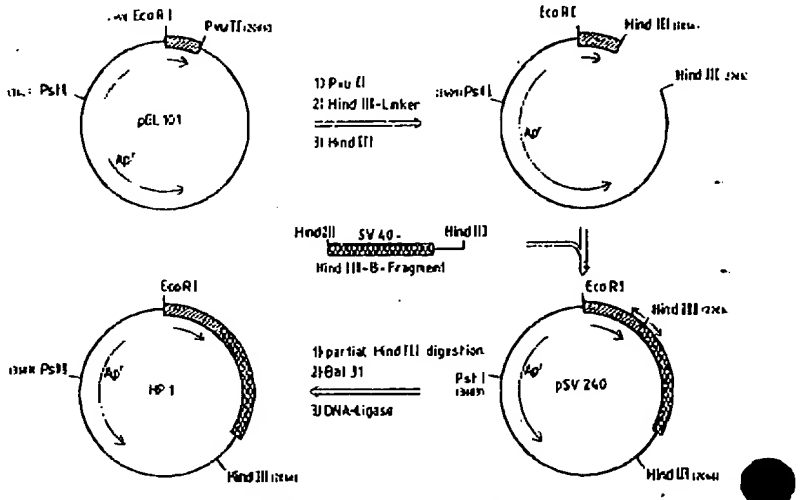


Fig. 7-60. Construction of an expression vector for SV40 t-antigen.
The *Pvu* II site of pGL101 (Fig. 7-10) is converted into a *Hind* III site by using *Hind* III linkers, which all subsequent insertion of the SV40 DNA fragment. In contrast to the construction shown in Fig. 7-59, the in *i.e.*, the t-antigen DNA fragment (cross-hatched bar), as well as the S/D sequence within the portable *lac* pr (stippled bar) are shortened by exonuclease treatment (indicated by arrows extending from the *Hind* III pSV240). Numbers in brackets are pBR322 co-ordinates. Directions of transcription are indicated by arrow the plasmid circle.

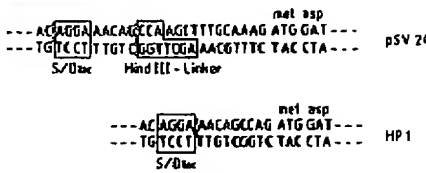


Fig. 7-61. Structure of hybrid ribosomal binding sites comprising the S/D sequences of the *lac* operon and the start codon of SV40 t-antigen.
S/D sequences and a part of the *Hind* III linker in pSV240 are boxed. In contrast to pSV240, 11 base pairs of the 5′ untranslated region of the SV40 t-antigen are missing in **HP1**; the distance between S/D sequence and ATG codon is therefore reduced to nine base pairs. The expression of t-antigen in **HP1** is 40-fold higher than in pSV240 (*cf.* also Table 7.2 and Fig. 7-60).

plasmid pSV240 was used to shorten the distance between the S/D sequence and the ATG codon by *Bal*31 treatment (Fig. 7-61). The properties of different clones, in particular their activity with respect to production of t-antigen, and the secondary structures of the ribosome binding sites are summarised in Table 7-2 and discussed in Section 7-2.

In principle, this procedure can also be used for any other gene to be expressed from the *lac* promoter (Fig. 7-62) (Guarente *et al.*, 1980a). A cDNA copy of the desired gene is first cloned in pBR322. A cleavage site is then introduced in the vicinity of the 5′ end by using a synthetic linker with a recognition site which does not occur in the cDNA itself. In the example shown in Fig. 7-62, a *Bam* HI linker was used. Other recommended and suitable cleavage sites are situated clockwise from the single *Eco* RI site in pBR322, *i.e.*, *Cla* I, *Hind* III, *Eco* RI, *Bam* HI (*cf.* example in Fig. 7-62) or *Sal* I. The resulting plasmid is opened at the desired site by partial digestion and the DNA is treated with exonuclease III before the promoter fragment is annealed. This cloning approach is simple because there is an efficient screen for those plasmids which contain the *lac* promoter region. The operator sequence on the insertion titrates the repressor molecules within the bacterial cells, the chromosomal β-galactosidase is expressed constitutively, and blue colonies appear on Xgal medium.

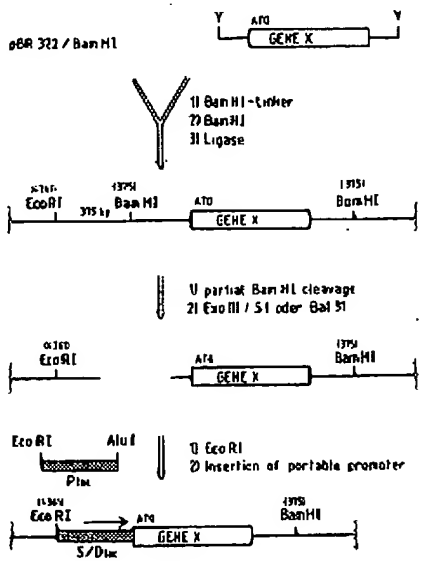


Fig. 7-62. General approach for the construction of expression vectors with foreign genes expressed under the control of a portable *lac* promoter.
Shown are the reactions to be carried out for bringing a hypothetical gene (X) under the control of a portable promoter. See text for details. Numbers in parentheses are pBR322 co-ordinates (Guarente *et al.*, 1980a).

One disadvantage of this procedure is that constructions which maximally express the inserted gene are not immediately identified. It is necessary to carry out functional or immunological tests for the expression of the desired protein with each clone. Since this may be comparatively laborious, a method was developed to allow identification of not only insertions of the *lac* operator/promoter region, but also optimal positioning of the hybrid ribosome binding site (Guarente *et al.*, 1980b). This technique exploits a particular property of the enzyme β-galactosidase. Large carboxy-terminal peptide fragments are enzymatically active irrespective of the nature of the *N*-terminal end. This fact was first observed with fusions of *lac* repressor and β-galactosidase (*lacI/lacZ*) (Müller-Hill and Kania, 1974). In order to generate such fused genes, a bacterial strain was constructed which contained a *lacI* gene promoter mutation (i^q) and an ochre mutation in *lacZ* (*UI18*). The i^q mutation causes an overproduction of repressor protein, and the ochre mutation terminates transcription in *lacZ* at the position of the seventeenth amino acid so that a truncated, enzymatically inactive β-galactosidase of sixteen amino acids is synthesised (7-63). A selection for revertants expressing galactosidase activity led to gene fusions lacked the termination signal for *lacI*, promoter region, and variable parts of terminal region of *lacZ*. This indicated th *N*-terminal part of β-galactosidase co replaced by parts of the *lac* repressor v influencing the enzymatic activity of β-gal dase. This important observation, which the basis of α-complementation, has late applied to fusions of *lacZ* with other prok genes.

It is important to realise that this princi also be applied to *N*-terminal regions of e otic genes (Guarente *et al.*, 1980b). B contains a large *C*-terminal part of (*lacZ′*), but this region is not expressed, si corresponding promoter regions are abse shown in Fig. 7-64, the 5′ terminal part o gene X, can be fused with the *lacZ* region plasmid in such a way that a functional r frame is generated. In the case of a euk

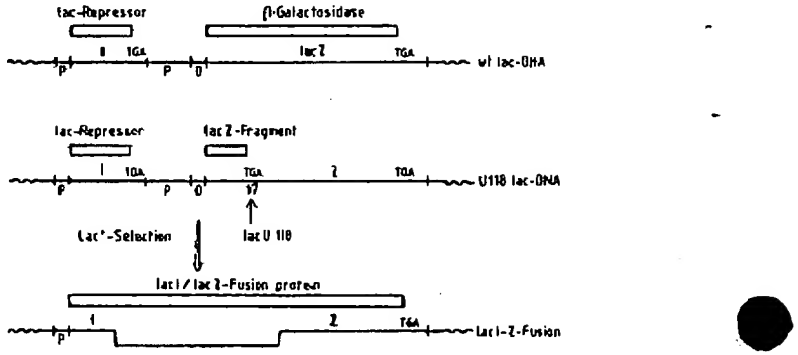
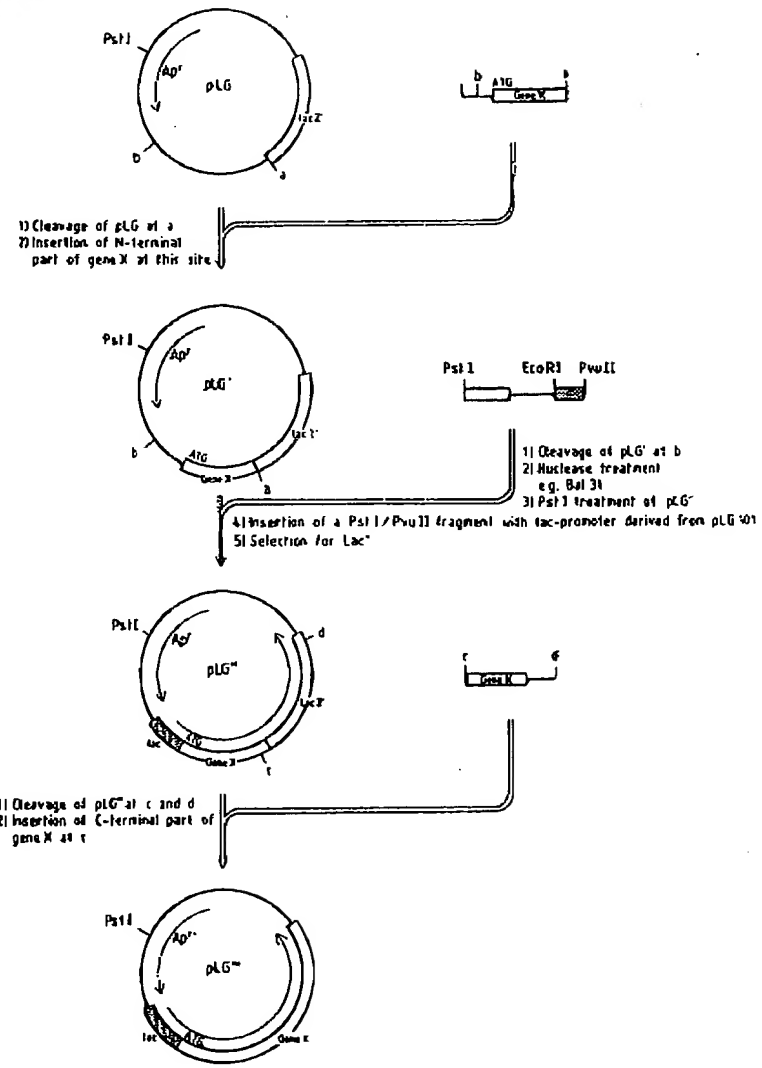


Fig. 7-63. Construction of *lacI-lacZ* fusions.
The top line shows a part of the *E. coli* chromosome with the arrangement of genes for *lac* repressor and w β-galactosidase. Mutation *UI18lac* is characterised by a stop codon at position 17 of β-galactosidase (*lac lacI-lacZ* fusion bypasses the stop codon, TGA, of the *lac* repressor gene (*lacI*), and also the nonsense c *lacZU118*. Proteins encoded by the DNA regions in question are shown as open bars above the maps (M0 and Kania, 1974).



β-Galactosidase	---TAACAATTCCACACAGGAAACAGCTATG---
SV40-t-Ag (pTR436)	---TAACAATTTCACACAGGAAACAGAAAGATG---
β-Globin (pLG 302-7)	---TAACAATTTCACACAGGAAACAGACAGAATG---
pre FIF (pLG 1141	---TAACAATTTCACACAGGAAACAGACATG---
FIF (pLG 117)	---TAACAATTTCACACAGGAAACAGCCATG---
	S/D box

Fig. 7-65. Ribosomal binding sites in different expression vectors.
All structures contain the same S/D sequence also found in the *lac* system (boxed). The point of tran sequences specific for various eukaryotic genes is indicated by a vertical line (Guarente *et al.*, 1980a).

◄- - **Fig. 7-64.** General approach for constructing optimally expressing clones in the *lac* system.
The 5' terminal part of a gene X to be expressed is introduced into a restriction site "a" of a plasmid containing the 3' terminal portion of the *lacZ* gene. Plasmid pLG' is then opened at "b", modified by nuclease treatment and ligated with a *lac* promoter fragment (obtained, for example, from pGL101 Fig. 7-10) which is flanked by *Pst* I and *Pvu* II sites. Maximally expressing clones are identified by selection for Lac$^+$ in a growth medium containing Xgal. The *lacZ'* region in these clones is then replaced by the 3' terminal portions of X by using site "c", which restores gene X (Guarente *et al.*, 1980).

gene this fusion in plasmid pLG' is
transcribed nor translated, since no pr
signals are available. A combination of res
enzyme digestions and nuclease treatmen
lar to those described above (Fig. 7-62)
used to insert a portable promoter in fron
first ATG of gene X. If the resulti
binding site in plasmid pLG'' has
structure, it will direct the synthesis of
matically active gene X/β-galactosidase
protein. The enzymatic activity of this p
which contains an *N*-terminal portion of
protein, is easily identified on Xgal plat
transformation of Lac-negative bacteria.
the *lacZ* part of the hybrid gene in the pla
replaced by the 3' terminal DNA fragn
gene X in order to obtain the entire gene
correct configuration (pLG'''). A compar
β-galactosidase activities of the fusion p
(X-*lacZ'*) obtained from pLG'' with the
of functional X protein in pLG''' shows tha
values correlate quite well. A clone expres
β-galactosidase fusion protein generally w
duce the intact eukaryotic gene product
Good producers can therefore be identifie
on the basis of β-galactosidase activity ex
from gene X/β-galactosidase fusions. At le
eukaryotic genes have been expressed
technique: small t-antigen of SV
globin, and human fibroblast inte
with and without its pre-sequence). Th

Benno Müller-Hill,
Cologne, FRG

Table 7-5. Expression yields of IGF-1/*lacZ* fusion proteins from different mutated plasmid constructions.

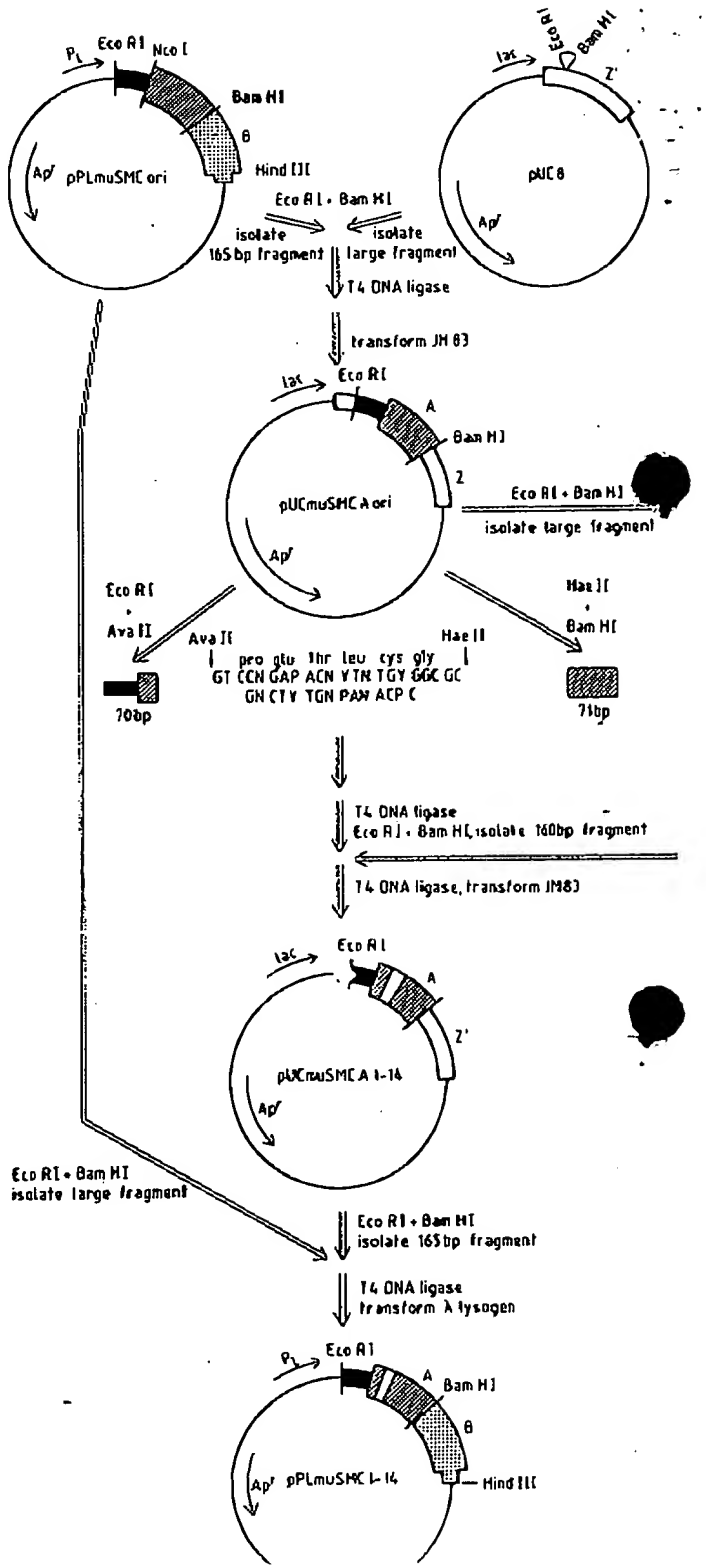
Plasmid	2 Pro	3 Glu	4 Thr	5 Leu	6 Cys	β-gal activity units/cell (JM83)	SMC ng/10^7 cells (HB101)
original sequence	x	x	x	x x	x		
pUCmuSMCA$_{ori}$	CCA	GAA	ACC	CTG	TGC	0.4	1.4
blue colonies							
pUCmuSMCA 1	CCC	GAA	ACT	CTG	TGT	3.1	33
2	CCT	GAA	ACT	TTG	TGC	2.6	45
3	CCA	GAG	ACG	TTG	TGC	0.9	35
4	CCA	GAG	ACG	TTG	TGT	0.9	43
5	CCT	GAA	ACT	TTG	TGT	2.9	33
6	CCT	GAG	ACG	TTG	TGT	1.2	58
7	CCG	GAA	ACG	TTA	TGT	1.9	50
8	CCG	GAA	ACA	TTG	TGT	1.2	65
9	CCA	GAA	ACG	TTG	TGT	1.1	32
10	CCT	GAG	ACT	CTA	TGT	2.3	22
white colonies							
pUCmuSMCA 11	CCC	GAA	ACC	CTC	TGT	<0.1	0.10
12	CCT	GAA	ACC	CTC	TGT	<0.1	0.11
13	CCG	GAA	ACC	CTC	TGT	<0.1	0.10
14	CCA	GAA	ACC	CTC	TGT	<0.1	0.09

The table relates to plasmid constructions described in Figs. 7–66 and 7–67. SMC stands for somatomedin-C, a 70 amino acid protein found in human serum, also known as insulin-like growth factor I (IGF-I). Numbers 1 to 14 refer to fourteen plasmid colonies, ten of which displayed a blue and four of which displayed a white phenotype following plating on JM83 cells in the presence of Xgal plus ampicillin. x indicates the positions of mutations introduced in codons 2 to 6 of the IGF-I gene (Buell *et al.*, 1985).

tures of the respective ribosome binding sites are shown in Fig. 7-65. At equilibrium the corresponding bacteria synthesise between 5 000 and 15 000 molecules of the desired protein per cell. In individual cases the yields may be lower since the various proteins may differ in their stability within the host bacteria (Guarente *et al.*, 1980a) (see Section 7.5).

This concept can also be applied to the pUC family of vectors. In order to optimise expression of the IGF-I gene in *E. coli*, Buell *et al.* (1985) inserted a 165 bp fragment containing a ribosome binding site and the coding region of the first 33 amino acids of IGF-I into the 65 bp polylinker region of pUC8 (Fig. 7-67). Expression in this construction was under the control of the *lac* promoter, while translation could initiate either at the *lacZ* gene or at the IGF-I gene start codon. However, since translation from the *lacZ* AUG would quickly encounter a stop codon (Fig. 7-67),

Fig. 7-66. Construction of IGF-*UlacZ* fusion vectors for improved expression of the IGF-I protein. Vector pPLmuSMCori contains a synthetic IGF-I gene (parts "A" and "B", hatched and stippled bars) preceded by a 66 bp fragment derived from bacteriophage mu (see Fig. 7-67) which provides the S/D sequence (black bar). The construction is driven by the λ P_L promoter, but results in only low level expression of the desired protein. To improve expression, the *N*-terminal part "A" of the IGF-I gene is cloned into pUC8 to yield pUCmuSMCAori. An *Ava*II-*Hae*III fragment, (G/G(AorT)CC) (PuGCGC/Py), coding for amino acids 2-8 of IGF-I, is replaced by a synthetic mixture of DNA fragments containing all possible base substitutions which retain the amino acid sequence. The *Eco* RI-*Bam* HI fragments from plasmids isolated from blue colonies were isolated and reconstructed into pPLmuSMC1-14, as indicated. The open bar sector within part "A" of the IGF-I gene in plasmids pUCmuSMCA1-14 and pPLmuSMC1-14 represents the synthetic fragment. N = one of the four possible bases, P = purines, and Y = pyrimidines (Buell *et al.*, 1985).



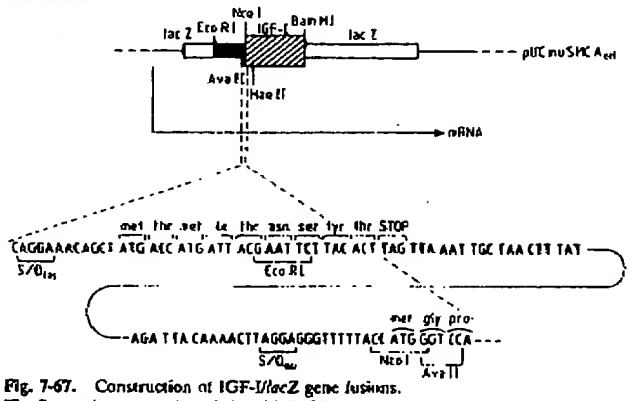


Fig. 7-67. Construction of IGF-I/*lacZ* gene fusions.
The figure shows a section of plasmid pUC8 with an insertion between its *Eco* RI-*Bam* HI polylinker sites within the *lacZ* α-peptide (open bar). The insertion comprises a 66 bp *Eco* RI-*Nco* I fragment containing the S/D sequence from the *nerI* gene of bacteriophage mu (Gray *et al.*, 1984) (black bar), and a 98 bp long *Nco* I-*Bam* HI fragment with the first half (part "A" in Fig. 7-66) of the coding region of the human IGF-I gene (hatched bar). Transcription in this construction starts at the *lacZ* promoter present in the pUC8 portion of the vector, and covers the *lacZ* portion and the insert. Translation from this message can initiate both at the *lacZ* ribosomal binding site (S/D_{lac}) and the ribosomal binding site (S/D_{mu}) of the insert derived from bacteriophage mu. Ribosomes initiating at the *lacZ* ribosomal binding site will soon encounter a stop codon, while the second translation product continues through the IGF-I coding region into the *lacZ* portion of pUC8. This construct yields only white plaques on *E. coli* strain JM83. Mutations introduced into the coding region within an *Ava* II-*Hae* II fragment (*cf.* Fig. 7-66) result in a blue phenotype, indicating expression of the IGF-I/*lacZ* fusion peptide (Buell *et al.*, 1985).

the only protein formed was derived from a fusion between the IGF-I portion and the distal *lacZ* gene region (Fig. 7-67). Transfection of a construction containing genuine IGF-I sequences into JM83 yielded only white colonies, indicating little or no β-galactosidase activity. In order to increase expression, a large number of mutants were generated by synthesising a mixture of oligonucleotides that included all the 256 possible sequences encoding amino acids 2-6 of IGF-I (Fig. 7-66). After re-insertion into the proper position in the fusion and transformation of *E. coli* strain JM83, approximately 500 out of 5 000 colonies were pale blue. The best of these, after reconstruction of the whole IGF-I gene, produced more than 20 times more IGF-I than did the wild-type construction (Table 7-5). These mutations affected a secondary stem structure around a ribosome binding site region and the increased β-galactosidase activity in the mutants thus confirms some of the conclusions mentioned in Section 7-2.

7.4.2.2 *The trp System*

As in the *lac* system, the regulatory sequences of the *trp* operon can be used to create hybrid ribosome binding sites. In the *trp* system, the site of transcription initiation and the start codon of the TrpE protein are separated by 162 bp known as the leader sequence (Fig. 7-14). This region codes for a peptide which is fourteen amino acids in length and plays a decisive role in the control of the *trp* operon. A *Taq* I site is situated between the corresponding S/D sequence and the ATG codon, allowing both parts of the ribosome binding site to be separated from each other. The ATG of the leader peptide can therefore be

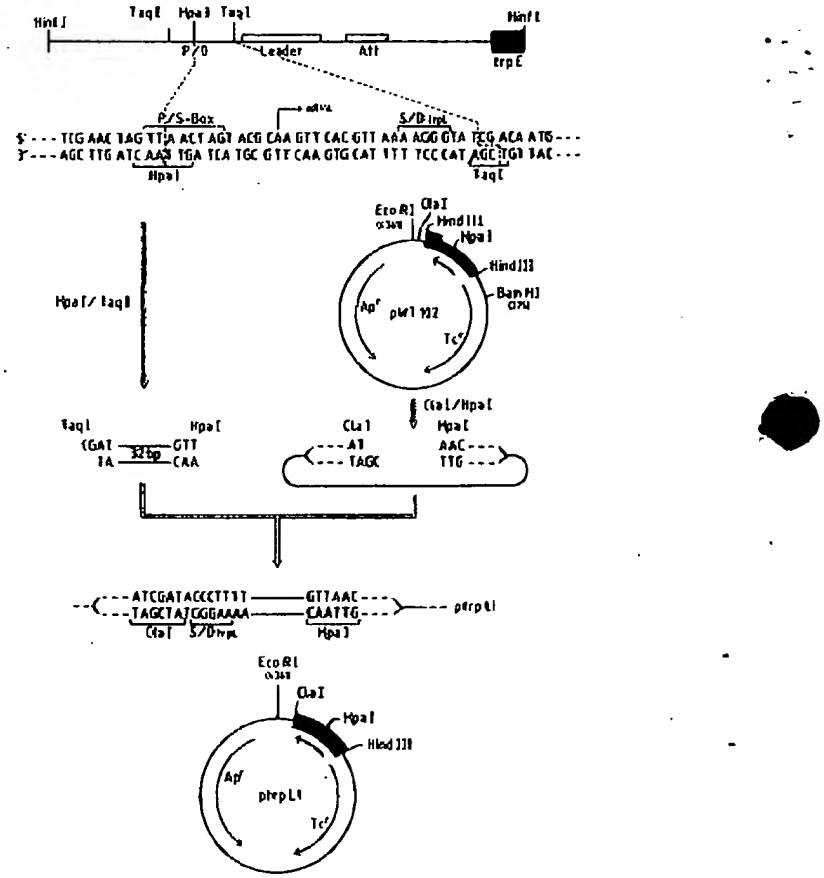


Fig. 7-68. Construction of an expression vector with *trp* regulatory sequences.
The top shows a portion of a *Hinf* I DNA fragment with *trp* regulatory sequences (*cf.* also Fig. 7-15) between a *Hpa* I and a *Taq* I site contains the Pribnow-Schaller box (P/S) and the S/D sequence of the *trp* peptide (S/D_{trpL}). This 32 bp *Hpa* I-*Taq* I fragment is inserted into vector pWT102 opened with *Hpa* I and pWT102 and pWT101 (Fig. 7-18) are identical, but contain the *trp* insertion in opposite directions (arrows) unexpectedly, pWT102 shows a considerable tetracycline resistance, which is presumably due to the presence cryptic promoter in the *trp Hinf* I fragment. By a fortunate coincidence, the *Taq* I site of the insertion contains pair, and this regenerates the *Cla* I site. Vector ptrpL1 lacks the *trp* coding regions, and the *Cla* I site can there used directly for cloning of foreign DNA (Edman *et al.*, 1981).

replaced by the ATG of a eukaryotic gene.

A suitable plasmid which contains the entire *trp* promoter region, the S/D sequence for the leader peptide, but no other parts of the *trp* operon was constructed from plasmid pWT102 (Fig. 7-18) (Edman *et al.*, 1981). Digestion with a combination of *Hpa* I (GTT/AAC) and *Taq* I (T/CGA) yielded a 32 bp fragment which contained the desired S/D sequence and the initiation site for transcription (Fig. 7-68). In a parallel experiment all *trp*-specific regions upstream of the *Hpa* I site were removed from the same plasmid, pWT102, by digestion with *Hpa* I and *Cla* I and replaced by the short *Hpa* I-*Taq* I fragment. This construction presented no problems since the tetrameric *Taq* I recognition sequence (T/CGA) is part of the hexameric recognition sequence of *Cla* I (AT/CGAT), and since both enzymes cleave in the same pattern with protruding 5'-CG termini. Since there was an A/T pair in the immediate vicinity of the original *Taq* I site, the *Cla* I site was regenerated. The resulting plasmid, ptrpL1, was opened at its unique *Cla* I site to allow insertion of a foreign gene in the immediate vicinity of the S/D sequence. The *Cla* I site is particularly well suited for this purpose since it can accomodate various DNA fragments with protruding 5'-CG ends, such as those obtained by *Hpa* II (C/CGG), *Taq* I (T/CGA) and *Acc* I (GT/CGAC) cleavage. This strategy was tested and employed for cloning and expressing hepatitis B virus core antigen (HBcAg). The gene for this protein, which is 183 amino acids in length, was obtained from a suitable plasmid by *Hha* I cleavage (Fig. 7-69). The start codon for HBcAg on a *Hha* I fragment of 1005 bp was 15 bp away from one of the molecular ends of this fragment. The distance between the ATG and the S/D sequence would have been too long, and therefore the usual modifications were carried out as described above. The DNA was first treated with exonuclease III to remove approximately ten base pairs and then made blunt-ended with S1 nuclease before *Bam* HI linkers were added. The commercially available decameric linkers do not only contain a

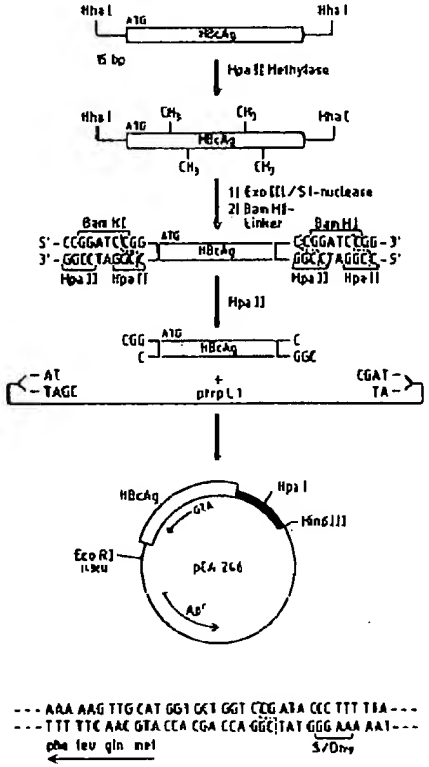


Fig. 7-69. Application of expression vector ptrpL1. A *Hha* I DNA fragment containing the coding sequence for HBcAg was modified by a combined exonuclease III S1 nuclease treatment. Following addition of *Bam* HI linkers, the construct was digested with *Hpa* II in preparation for cloning into the *Cla* I site of ptrpL1. The sequence around the ribosomal binding site at the bottom shows that the distance between S/D sequence and ATG codon is 16 bp. The direction of transcription of the HBcAg sequences is indicated by an arrow.

Bam HI site but also two *Hpa* II sites. A subsequent *Hpa* II digestion completely removes the *Bam* HI recognition site and creates 5' overhanging ends which are compatible with *Cla* I ends. Since the tetrameric *Hpa* II recognition site occurs

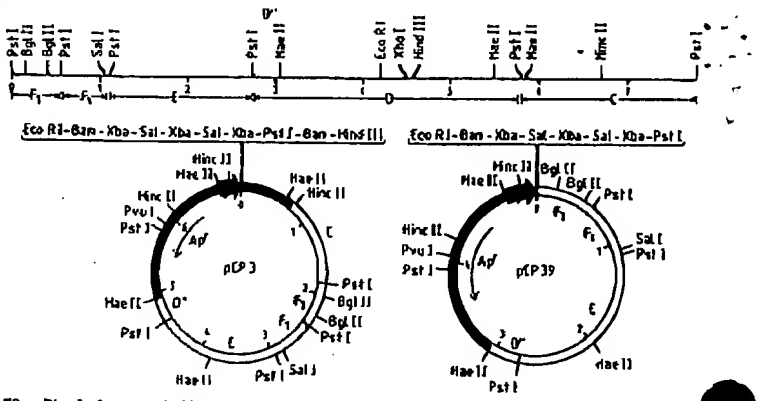


Fig. 7-70. Physical maps of pKN402 and its derivatives. Plasmid pKN402 (shown on top in a linear presentation) is a miniderivative of a temperature-sensitive replication mutant of plasmid R1drd19. Plasmids pCP3 and pCP39 are derivatives of pKN402, which contain *Pst* I-F2 fragment required for thermoinducible runaway replication as well as a selectable marker unique pKN402. Capital letters indicate the *Pst* I fragments of pKN402. Plasmid pCP39 lacks a 1790 bp *Pst* I fragment present in pCP3, which represents part of the pKN402 *Pst* I-C fragment. The ampicillin resistance gene (blac and the λ P_L promoter (black bars with arrow) are derived from the pPLa series of plasmids described in Fig and 7-27 (Remaut *et al.*, 1983).

quite frequently, the DNA fragment must be protected by treatment with *Hpa* II methylase prior to *Hpa* II digestion. In principle, this strategy can be used for any other gene. A disadvantage is that it does not directly allow selection of or quick screening for maximally expressing clones. In our example (Fig. 7-69) screening of a large number of transformants yielded the expression vector pCA246, which produces up to 10% of the newly synthesised protein as HBcAg after induction with 3-β-indolylacrylic acid.

7.4.2.3 The λ P_L System

The strong λ P_L promoter has been particularly useful for the high-level expression of proteins detrimental to an *E. coli* cell. In an elegant and most efficient application it is employed in a two plasmid system (Remaut *et al.*, 1983) which also exploits the temperature-sensitive runaway replication phenomenon alluded to earlier (Se 4.1.1). One plasmid component is derived plasmid pKN402, a 7.8 kb mini-derivative runaway replication mutant of plasmid R1c (Fig. 7-70). This plasmid contains both the perature-sensitive replicon and the λ P_L pro er; the latter lies upstream from a polylinker, which the desired gene can be inserted. Ex sion of the P_L promoter from such a construc be regulated by the *cI* gene product encoded single chromosomal gene copy or, even bette a *cI* gene on a compatible multicopy plas Such a plasmid, pcI857, is described in Se 4.1.5. It confers kanamycin resistance, *cI857* allele of the λ repressor, and is co with the replicon of pKN402 and its derivat *E. coli* cells transformed with both the run replication P_L vector and the pcI857 ve contain approximately 30-50 copies of each ve at 28 °C. At this temperature the active c repressor acts *in trans* to prevent any transcrip

from the P_L promoter on the other plasmid. A shift to higher temperature (42 °C) leads to two events, a ten- to twentyfold amplification of the runaway replication vector copy number, and a simultaneous derepression of the P_L promoter due to inactivation of the cI857 repressor at 42 °C. This two-plasmid expression system was tested with the T4-derived DNA ligase gene, the expression of which could be induced to levels up to 25% of the total cellular protein. It is effective in many *E. coli* strains and has also proved successful for the expression of the human IGF-I protein (Buell *et al.*, 1985).

7.4.2.4 Synthetic Ribosome Binding Sites

The hybrid ribosome binding sites discussed in Sections 7.4.2.1 and 7.4.2.2 are not necessarily optimal for ribosome binding, and hence for efficient translation (*cf.* also Section 7.2). These binding sites contain naturally occurring S/D sequences which frequently show a relatively low degree of homology with the sequence of the 3′ end of 16S ribosomal RNA. In the *lac* system it is only four and in the *trp* leader peptide S/D sequence only three bases which show this homology at all. It was postulated (Jay *et al.*, 1981) that ribosome binding, and hence initiation of protein biosynthesis, would be much more efficient if these regions of homology could be extended. A DNA oligomer containing an S/D sequence of nine base pairs and an additional sequence,

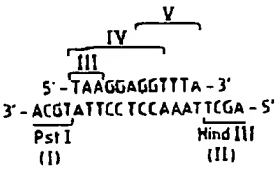


Fig. 7-71. Structure of a synthetic linker with *Pst* I (I) and *Hind* III (II) ends, coding for a stop codon (III), an S/D sequence (IV), and the GGTTTA sequence. (Jay *et al.*, 1981).

5′-GGTTTAA-3′, which is important for binding ribosomal proteins (Fig. 7-71; *cf.* also Fig. 7-37; also Jay *et al.*, 1982) therefore was synthesised chemically. The entire synthetic ribosome binding site consists of two oligonucleotides of twelve and twenty bases, respectively. The left-hand 3′ protruding end contains a sequence which allows ligation with a *Pst* I site (I), the right-hand 5′ protruding end a *Hind* III site (II). A TAA stop codon (III) within this linker molecule is in phase with β-lactamase (see below); in the inner part of this linker lie the S/D sequence of nine bases (IV) and the sequence GGTTTAA (V). Since the linker is asymmetrical it is more universally applicable than conventional symmetrical linkers.

As shown in the example in Fig. 7-72, this linker is positioned at a correct distance in front of the start codon of a gene to be expressed, and inserted together with this gene X into the *Pst* I site within the β-lactamase gene of pBR322 (*cf.* also Fig. 4.1-11). In a bacterial cell, transcription initiates at the promoter of the β-lactamase gene to yield a hybrid mRNA containing the β-lactamase component and sequences of gene X

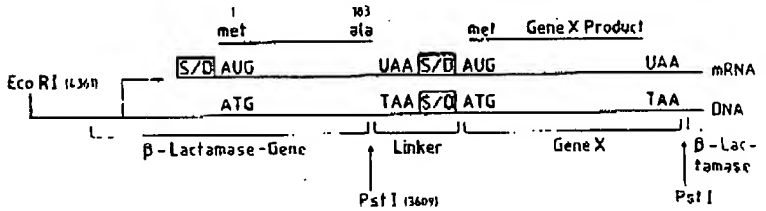


Fig. 7-72. Use of synthetic ribosomal binding sites for the construction of expression vectors. The linker carries a stop codon and a consensus S/D sequence. Although only one hybrid mRNA is transcribed, two proteins are synthesised, one of which is a fragment of β-lactamase with amino acids 1-183; the other is the gene X product with an *N*-terminal methionine residue. Numbers in parenthesis are pBR322 co-ordinates.